HW 6 Movie Recommendation

學號：r06323011 系級：經濟碩一 姓名：葉政維

1. 請比較有無 normalization 的差別。並說明如何 normalize.

2. 比較不同的 embedding dimension 的結果。

題 1、2 皆使用 matrix factorization。除了控制 user、movie 的 bias 外，也將其他 user、movie-specific 的特質加入 bias 中。此外，由於擔心因為收斂太快而沒有辦法在各 epochs 中挑到最佳模型，因此使用比較大的 batch_size (2000)。Normalization 的方法為：將 rating 減去平均值再除上標準差。（註：所有 prediction 的結果若超過上限 5 或下限 1，則重設為 5、1）

有無 Normalization 在 private score (RMSE) 的表現如下：

Emb Size	10	50	100	150	200
Base	0.8610	0.8502	0.8487	0.8485	0.8471
Normalize	0.8643	0.8574	0.8545	0.8536	0.8522

不做 Normalization 以及較大的 embedding dimension 會有比較好的表現，儘管如此有 Normalization 時收斂得較快。較大的 embedding dimension 能更仔細的描述人、電影的潛藏特質，有助於模型表現。Normalization 的結果則較不合預期，猜測是因為估計 mean、std 時的 variation 帶來的影響。

3. 比較有無 bias 的結果。

5. 試著使用除了 rating 以外的 feature，並說明你的作法和結果，結果好壞不會影響評分。

題 3、5 使用的模型為：dimension=185 且不進行 normalize，並進一步比較 {是否有 id bias} x {是否使用 feature bias} 所組合出的四種結果。使用的 feature 包含：年齡、性別、職業、電影年份（較早期的 group 在一起）以及電影種類。使用方法是將這些變數依類別 embed 成 1 維的 bias term，並和 MF 的內積項加在一起（如同 id bias）。電影種類因為可能有多種類別，所以每個樣本點的電影種類先被 pad 成長度為 6 的 array（不足的補「Empty」），個別 embed 成 1 維的 bias term 後再加總。

「都沒有」、「Id bias」、「feature bias」、「皆有」的 private score 分別為：0.8516、0.8515、0.8479、0.8470。有加入 id bias 以及 feature bias 的結果都有進步，但是 feature bias 帶來的進步比 Id bias 更大，猜測是因為有些 id 只在 rating data 中出現過幾次，甚至完全沒有，因此並沒有 well-trained。

4. 請試著將 movie 的 embedding 用 tsne 降維後，將 movie category 當作 label 來作圖。

將 movie 的 embedded vector (185 維) 利用 Tsne 降至 2 維，作圖後的一些觀察大致符合直覺，以下舉幾個例子作為比較

- 見下圖 1，Horror 正好填在 Drama 與 Comedy 的「空缺之間」，顯示 Horror 和其他兩類很不一樣。
- 見下圖 2，Adventure、Child、Musical 三類的分布非常靠近，顯示這三類是相近的。